Atmospheric Pollution Research 5 (2014) 696-708

Atmospheric Pollution Research

www.atmospolres.com

Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis

Madhavi Anushka Elangasinghe [1], Naresh Singhal [1], Kim N. Dirks [2], Jennifer A. Salmond [3]

[1] *Department of Civil and Environmental Engineering, The University of Auckland, Private Bag 92019, Auckland, New Zealand*
[2] *School of Population Health, The University of Auckland, Private Bag 92019, Auckland, New Zealand*
[3] *School of Environment, the University of Auckland, Private Bag 92019, Auckland, New Zealand*

ABSTRACT

Little attention is given to applying the artificial neural network (ANN) modeling technique to understand site–specific air pollution dispersion mechanisms, the order of importance of meteorological variables in determining concentrations as well as the important time scales that influence emission patterns. In this paper, we propose a methodology for extracting the key information from routinely–available meteorological parameters and the emission pattern of sources present throughout the year (e.g. traffic emissions) to build a reliable and physically–based ANN air pollution forecasting tool. The methodology is tested by modeling NO_2 concentrations at a site near a major highway in Auckland, New Zealand. The basic model consists of an ANN model for predicting NO_2 concentrations using eight predictor variables: wind speed, wind direction, solar radiation, temperature, relative humidity, as well as "hour of the day", "day of the week" and "month of the year" representing the time variations in emissions according to their corresponding time scales. Of the three input optimization techniques explored in this study, namely a genetic algorithm, forward selection, and backward elimination, the genetic algorithm technique gave predictions resulting in the smallest mean absolute error. The nature of the internal nonlinear function of the trained genetically–optimized neural network model was then extracted based on the response of the model to perturbations to individual predictor variables through sensitivity analyses. A simplified model, based on the successive removal of the least significant meteorological predictor variables, was then developed until subsequent removal resulted in a significant decrease in model performance. The developed ANN model was found to outperform a linear regression model based on the same input parameters. The proposed approach illustrates how the ANN modeling technique can be used to identify the key meteorological variables required to adequately capture the temporal variability in air pollution concentrations for a specific scenario.

Keywords: *Air pollution modeling, artificial neural networks, sensitivity analysis, meteorology, NO_2 concentrations*

doi: 10.5094/APR.2014.079

Corresponding Author:
Madhavi Anushka Elangasinghe
☎ : *+64-9-3737599*
☏ : *+64-9-3737462*
✉ : *mela005@aucklanduni.ac.nz*
manushkae@yahoo.com

Article History:
Received: 11 December 2013
Revised: 23 May 2014
Accepted: 23 May 2014

1. Introduction

Concentrations of pollutants in the atmosphere are influenced by the strength of emission sources, chemical transformations and atmospheric conditions (Jiang et al., 2005). For near–surface emissions (such as vehicle–generated pollutants), the surface winds determine the transport and dispersion of pollutants, and the atmospheric stability determines the extent to which pollutants are dispersed vertically within the atmospheric boundary layer (Hewson, 1956). Solar radiation and temperatures are also important in nitrous oxide conversion chemistry (Jiang et al., 2005). Variations in emission strengths and the surface meteorology can be observed by monitoring over different averaging times (such as hourly, daily and monthly). The modeling of atmospheric pollutant concentrations typically involves the development of a functional relationship between concentrations and the above–mentioned controlling parameters. One approach is to use deterministic models, relying on the governing fluid dynamic and chemical transformation mechanisms to model this relationship, while statistical models use field measurements of emission rates, meteorological parameters and concentrations to develop a linear or non–linear function between the concentration and these predictor variables. Deterministic models are limited by their requirement for detailed knowledge of source parameters, the topographical structures in the immediate surroundings and the detailed meteorology. These data are not always available in practice. Purely statistical models, when adequately trained, may provide good predictions using routinely available data. However, they are limited by their inability to provide insight into dispersion mechanisms and hence may not be used in "what–if scenario analysis" with respect to changes in emission rates and meteorological conditions without performing a procedure such as a sensitivity analysis.

Most cities around the world have routine meteorological stations that measure basic meteorological parameters. Variables such as boundary layer depth and stability indices are not readily available in routine networks, even though they are known to be important in the dispersion processes of atmospheric pollutants (Hewson, 1956). Detailed information about emission source strengths is limited to a few cities. For the aforementioned reasons, more research is needed in developing air quality models that can capture adequately the variability in observed concentrations using limited knowledge of the values of meteorological and emission parameters. The ability of the ANN technique to capture the nonlinear behavior of complex atmospheric processes makes it a suitable tool for developing such models. Specifically, multilayer perceptron ANN models have been used widely in atmospheric sciences in recent years for prediction, function approximation and pattern classification (Gardner and Dorling, 1998).

Many recent studies have shown that ANN–based air pollution models perform better than other statistical techniques (Gardner and Dorling, 1999; Chelani et al., 2002; Grivas and Chaloulakou,

2006; Singh et al., 2012) and ANN–based air pollution forecasting models are being implemented in some cities (Jiang et al., 2004; Kurt and Oktay, 2010; Perez, 2012). An ANN model, trained for a particular site, can be used with confidence only for that site, as the local meteorological conditions and emission pattern that determine the processes controlling the pollutant behavior will vary between sites. Therefore, an ANN model has to be constructed and trained for each air pollution measurement site of the city when constructing a forecasting network.

Several studies have focused on improving ANN model performance using different input data classification techniques (Nagendra and Khare, 2006; Hrust et al., 2009; Kurt and Oktay, 2010; Cheng et al., 2012; Perez, 2012) and hybrid model optimization techniques (Grivas and Chaloulakou, 2006; Karatzas and Kaltsatos, 2007; Antanasijevic et al., 2013). However, very few recent studies have investigated its internal mechanism in order to understand the extent to which the modeled function identifies the relative contribution of these controlling emissions and meteorological parameters to the observed levels of concentrations (Singh et al., 2012; Yan Chan and Jian, 2013). This explicit knowledge can be used to construct simpler ANN models with practical importance and better generalized performance and hence increase the use of advanced ANN–based techniques for air pollution modeling. Some studies that use only basic meteorological parameters to model the concentration of several pollutants use the concentrations of other pollutants (for example, concentrations of PM_{10} and SO_2 for modeling NO_2) as predictor variables in the model (Singh et al., 2012). Such models have limited practical use, especially in forecasting, as they require measurements of other pollutants as the predictor variables. Time–lagged models (models that can forecast concentrations), for example, one hour ahead, three hours ahead or 24 hours ahead, are found to give reliable predictions (Gardner and Dorling, 1999; Karatzas and Kaltsatos, 2007) but are not very useful when large gaps in data are present due to equipment failure or calibration down time. Therefore, the main goal of this study is to demonstrate a methodology for extracting explicit knowledge on the relative contribution of different meteorological parameters on pollutant concentrations so that it may be used in the construction of a robust ANN model (based on pattern recognition) that is useful in situations where one has to rely only on a few meteorological predictors to model pollutant concentrations. It can also be used to identify the key meteorological parameters that should be measured with a greater sensitivity so as to be able to make accurate model predictions.

Nitrogen oxides are emitted into the atmosphere primarily from vehicle exhaust as nitric oxide and nitrogen dioxide. The nitric oxide reacts with ozone to form nitrogen dioxide (Gardner and Dorling, 1999). As in most major cities, vehicular emissions are the major source of nitrogen oxide emissions in the study area. Epidemiological studies have revealed associations between NO_2 concentrations and daily mortality from respiratory and cardiovascular causes and hospital admissions for respiratory conditions (Burnett et al., 1998). A study conducted in the UK has found that emission control measures have not resulted in a significant decline in atmospheric NO_2 concentrations (Carslaw et al., 2011). Hence, the forecasting of NO_2 concentrations is important and is chosen for this study.

2. Methodology

2.1. Auckland case study–site description and data

The site selected for testing the proposed methodology is within a suburb in Auckland, located in a narrow isthmus in the North Island of New Zealand. The complex coastline and low–lying undulating topography result in complex surface wind flow patterns, especially in conditions of low synoptic wind flows when sea–land breezes dominate the surface wind. In this study, the emphasis has been placed on using a minimal set of meteorological predictors that are readily observed at almost all meteorological stations in New Zealand to ensure that the model is of practical use. The hourly average NO_2 concentration data and meteorological data have been obtained from an automatic monitoring station managed by the Auckland City Council, deployed in the corner of the school grounds of Westlake Girls' High School, Takapuna, a suburb located about 10 km north of Auckland City. The site is exposed to winds from all directions due to the open nature of the site. It is bounded to the east by Wairau Road and to the west by State Highway 1 (50 m from the highway). The houses in this area are mixed in age from 1960s construction, and 75% of them have chimneys. The working yard of Atlas Concrete, a concrete manufacturing company and the Wairau Industrial Park are located approximately 100–200 m away. Therefore, there is a complex mixture of local emissions from traffic, home heating and industrial sources. However, the main source of NO_2 is traffic and industrial in nature (Davy et al., 2009). The meteorological tower on site measures wind speed, wind direction, ambient temperature, relative humidity and solar radiation, all at a height of 10 m from ground level, consistent with the meteorological parameters commonly measured in automatic weather stations across New Zealand. The inlet of the NO_2 sampler is located at 3 m above the ground.

2.2. Model building and analysis of the contribution of different predictor variables

A schematic diagram of the methodology proposed for understanding the contribution of different predictor variables from the routine monitoring network using ANN–based models is presented in Figure 1. In Step I, the available data from the routine network are analyzed, as data visualization and exploration is considered an important initial step in statistical modeling (Samarasinghe, 2006). The emissions from a particular source are complicated by changes in the frequency of specific dispersion conditions, such as variations in meteorology between different periods of the year (Malby et al., 2013) and building an ANN model with carefully–selected inputs with an understanding of the influencing time scales and wind speed/direction interactions giving a physical basis to the model. Among the many methods available for visualizing ambient air pollution and meteorological data (Carslaw and Ropkins, 2012; Malby et al., 2013), multivariate polar plots, scatter plots, wind roses and time variation plots were used for this case study as they were able to capture the predominant features of the specific data set. However, many other techniques suggested in the literature could also be used on a case–by–case basis.

In Step II (Figure 1), the training, cross validation and testing data sets are defined. Other studies have identified that, for emission sources present continuously throughout the year (e.g. traffic emissions), a full year of data provides sufficient information to develop statistical models, ensuring that seasonal factors affecting concentrations are taken into account (Carslaw and Carslaw, 2007; Arhami et al., 2013). Therefore, one year of data is chosen for the training while two weeks are separated for testing to see if the model has correctly captured the variability in concentrations for the training year. The forecasting ability of the model is tested on the consecutive year. Therefore, under the assumption that the emission pattern does not change significantly from one year to the next, the proposed model can be used to forecast concentrations for the consecutive year by providing values of new predictor variables. Data are then randomized (shuffled or permutations made on the spread sheet) to ensure the robustness of the trained network by providing normally distributed samples for training, cross validation and testing. The building and training of the network is then carried out using NeuroSolutions Version 6.27 (http://www.neurosolutions.com/) and normalized values of the inputs are introduced to the input nodes so that all variables fall in a small range, avoiding

discrepancies that lead to faulty interpretation by the model due to the large weights adopted by inputs with larger magnitudes (Samarasinghe, 2006). Detailed information about the step–by–step development of a neural network model is found in Principe et al. (1999) and Samarasinghe et al. (2006). After a large number of test runs, a multilayer perceptron, one hidden layer network with a Levenburg Marquardt back propagation algorithm having a hyperbolic tangent transfer function in the hidden layer and bias transfer function in the output layer, has been found to be the best topology. The step size, momentum rate and processing elements are optimized through genetic optimization (Samarasinghe, 2006).

In Step III (Figure 1), input parameter optimization is carried out to eliminate the irrelevant inputs from the network. Among several techniques suggested in the literature (Olden and Jackson, 2002; Olden et al., 2004), forward selection, backward elimination, genetic algorithm with sensitivity analysis techniques were applied in this study. These techniques are used in a few ecological modeling studies in the literature (Lek et al., 1996; Olden et al., 2004). Sensitivity analysis provides extra knowledge on the response of the network to changes in each of the meteorological and emission parameters. This is achieved by studying the response of the network by varying each predictor variable within a range in small steps while locking all other input parameters at their mean value and plotting the response of the model to perturbations to each predictor variable (Principe et al., 1999; Samarasinghe, 2006). With this knowledge, different models were constructed by removing different input parameters to understand the degree of importance of these parameters in explaining the variability in atmospheric pollutant concentrations.

In the final step, the results from Step III are analyzed and compared to identify the key meteorological variables required to adequately capture the temporal variability in air pollution concentrations for the specific scenario and to test the robustness of the ANN model to identify the key dispersion mechanisms.

2.3. Multi linear regression (MLR) models

To test if the model performance is biased towards the fact that the temporal pattern of NO_2 in 2010 is similar to that in 2011 and if the ANN model is capturing any non–linearity that is not picked up by a multi linear regression (MLR) model, three linear models were developed on the same data set. One MLR model was developed using inputs that represent time scales alone (month, day and hour), one model with time scales, wind speed and wind direction, and another model with time scales and all meteorological inputs. The results of the ANN model were compared against the results of these three models.

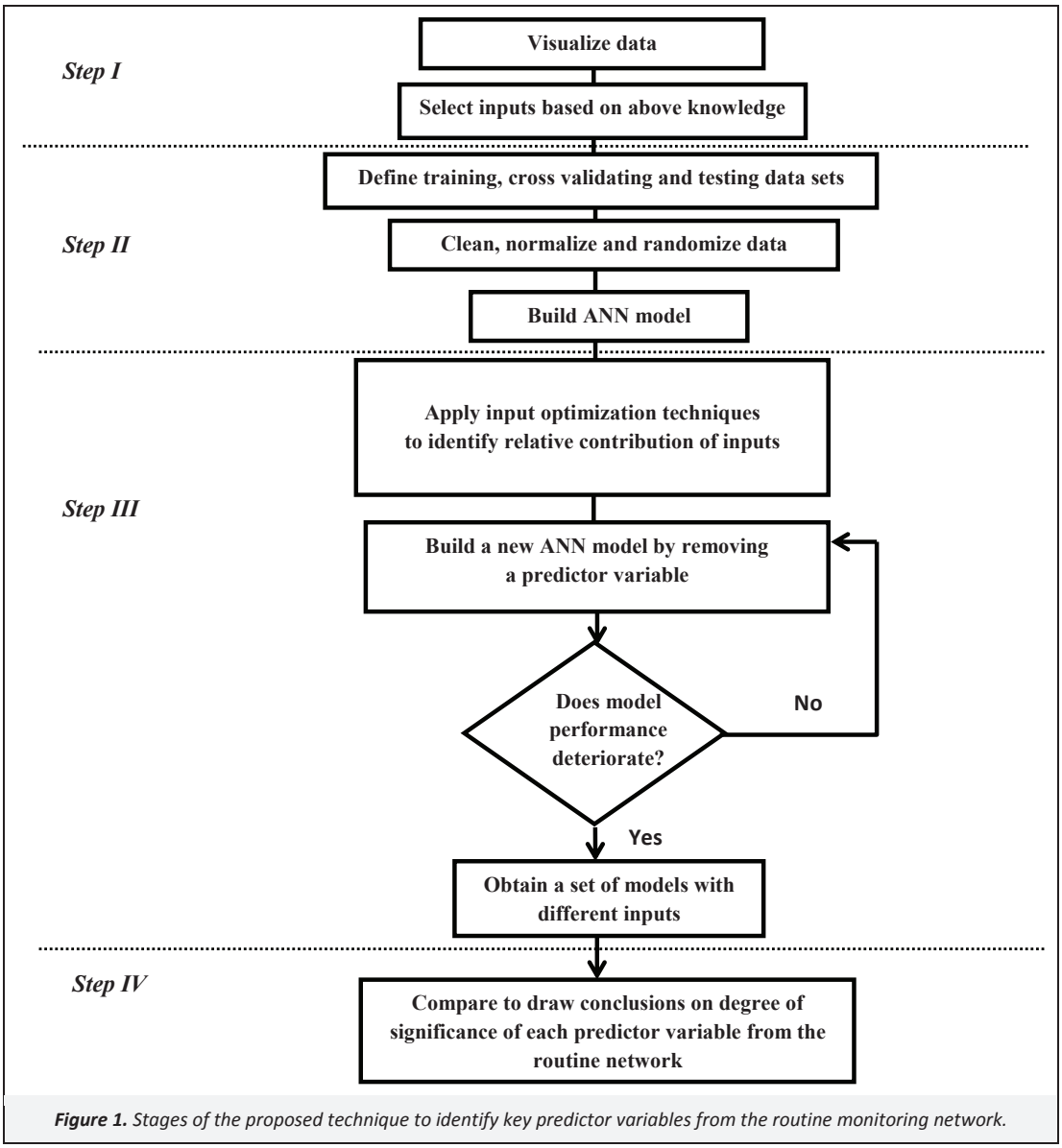

Figure 1. *Stages of the proposed technique to identify key predictor variables from the routine monitoring network.*

2.4. Model performance evaluation

The analysis of the performance of different ANN models developed in this study is based on a comparison of the model results for a test data set with actual observations. The following model performance statistics are used for comparing the model results, statistics commonly used in air quality model analysis (Carslaw, 2014). In the following equations, O_i denotes the i^{th} observed value, and P_i denotes the i^{th} predicted value for a total number of n observations.

A number of model evaluation parameters are considered. Namely, the fraction of predictions within a factor of two (FAC2) given by,

$$0.5 \leq \frac{P_i}{O_i} \leq 2.0 \quad (1)$$

The mean bias (MB) given by,

$$MB = \frac{1}{n}\sum_{i=1}^{n} P_i - O_i \quad (2)$$

The root mean squared error (RMSE) given by,

$$RMSE = \left(\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}\right)^{\frac{1}{2}} \quad (3)$$

The coefficient of determination (r^2) given by,

$$r^2 = \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{P_i - \bar{P}}{\sigma_P}\right)\left(\frac{O_i - \bar{O}}{\sigma_O}\right)\right]^2 \quad (4)$$

And the refined index of agreement (IA) (Willmott et al., 2012) given by,

$$IA = \begin{cases} 1 - \dfrac{\sum_{i=1}^{n}|P_i - O_i|}{c\sum_{i=1}^{n}|O_i - \bar{O}|}, when \\ \sum_{i=1}^{n}|P_i - O_i| \leq c\sum_{i=1}^{n}|O_i - \bar{O}| \; with \; c = 2 \\ \dfrac{c\sum_{i=1}^{n}|O_i - \bar{O}|}{\sum_{i=1}^{n}|P_i - \bar{O}|} - 1, when \\ \sum_{i=1}^{n}|P_i - O_i| > c\sum_{i=1}^{n}|O_i - \bar{O}| \end{cases} \quad (5)$$

3. Results and Discussion

3.1. Step 1: Visualization of data and input selection

The meteorological and air quality data sets for 2010 and 2011 were first used for the training and testing of the forecasting ability of the model as the first step in the modeling procedure. According to the statistical properties of variables given in Table 1, the concentration and meteorological parameters show similar distributions throughout the years. The meteorological parameters are highly variable, due to seasonal variations and the coastal nature of the site. As revealed by the wind rose diagram (Figure 2a), the prevailing winds are from the west, with significant frequencies also observed from the northeast and southwest during both years. High concentrations are observed in light to moderate wind conditions (Figure 2b), mainly when the winds are parallel to the highway (Figure 2b and 2c). Concentrations are higher when the site is downwind of the highway compared to when it is upwind (Figure 2b). Therefore, the highway seems to play a major role in elevated NO_2 concentrations, with a clear inverse relationship between NO_2 concentrations and wind speed.

Figure 3 illustrates how NO_2 concentrations are related to temperature, relative humidity, and solar radiation, as well as how these predictor variables are interrelated. At high values of solar radiation (during the day time), concentration is linearly related to both temperature (positively) and relative humidity (negatively). Relative humidity versus temperature shows a negative linear relationship. It also shows that NO_2 concentrations are negatively correlated with temperature and solar radiation and positively correlated with relative humidity. However, the scatter plots are too noisy to derive statistically significant linear relationships.

Hourly, daily and monthly variations in NO_2 concentrations are then analyzed. These are presented in Figure 4. The average hourly NO_2 concentration variations clearly show a diurnal variation with a morning peak around 8 am and an afternoon peak around 7.30 pm (Figure 4a). According to the monthly averages, the concentrations of NO_2 are high during the winter months of June, July and August due to elevated emission sources in conjunction with cold, calm weather. The concentrations are low during the summer months of December, January and February as a result of high winds and warm, dry weather (Figure 4b). When the average daily variation is analyzed, relatively high concentrations are observed during mid–weekdays compared to weekend days, while concentrations on Sunday are at their minimum (Figure 4c). Slightly lower average concentrations are observed in 2011 compared with 2010 (Figure 4a, 4b and 4c). It is also observed that the mean wind speed is 2.4 m s^{-1} in 2010 and 2.6 m s^{-1} in 2011. The low average concentrations in 2011 could have resulted from slightly higher wind speeds (or other meteorological conditions favoring effective dispersion) in 2011 or due to a reduction in emission sources.

The analysis provided above forms the basis for selecting input parameters to the ANN model. Since information on emission rates is scarce in many cities, it is not included as a model parameter. However, to make good model predictions, the emission pattern, normally unique in terms of its diurnal, weekly and monthly pattern, should be included in the model. This is achieved by including "hour of the day" (numbers from 0–23), "month of the year" (numbers from 1–12) and "day of the week" (numbers from 1–7, 1–Sunday to 7–Saturday) as inputs, along with other meteorological input parameters from the routine monitoring network, including wind speed (WS), wind direction (WD), temperature (T), relative humidity (RH) and solar radiation (SR). Hence, the final set of inputs selected for the modeling exercise consist of "hour of the day", "month of the year", "day of the week", temperature, relative humidity, wind speed, wind direction and solar radiation.

3.2. Step II: The ANN model

The chosen ANN model is represented graphically in Figure 5. The ANN model has been trained and genetically optimized on one year of data (2010) while two weeks are separated for testing for the same year. The optimized weights of the trained ANN model for this case are presented in Appendix B. The forecasting ability of the model for the following year is tested on the set of 2011 data after training on 2010 data.

3.3. Step III (a): Optimization of inputs

Selecting the best subset of inputs is the next step in model building as the main goal of this study is to achieve model simplicity for better generalized performance when making forecasts. Tables 2 and 3 summarize the results of the three standard input optimization techniques applied to this case study.

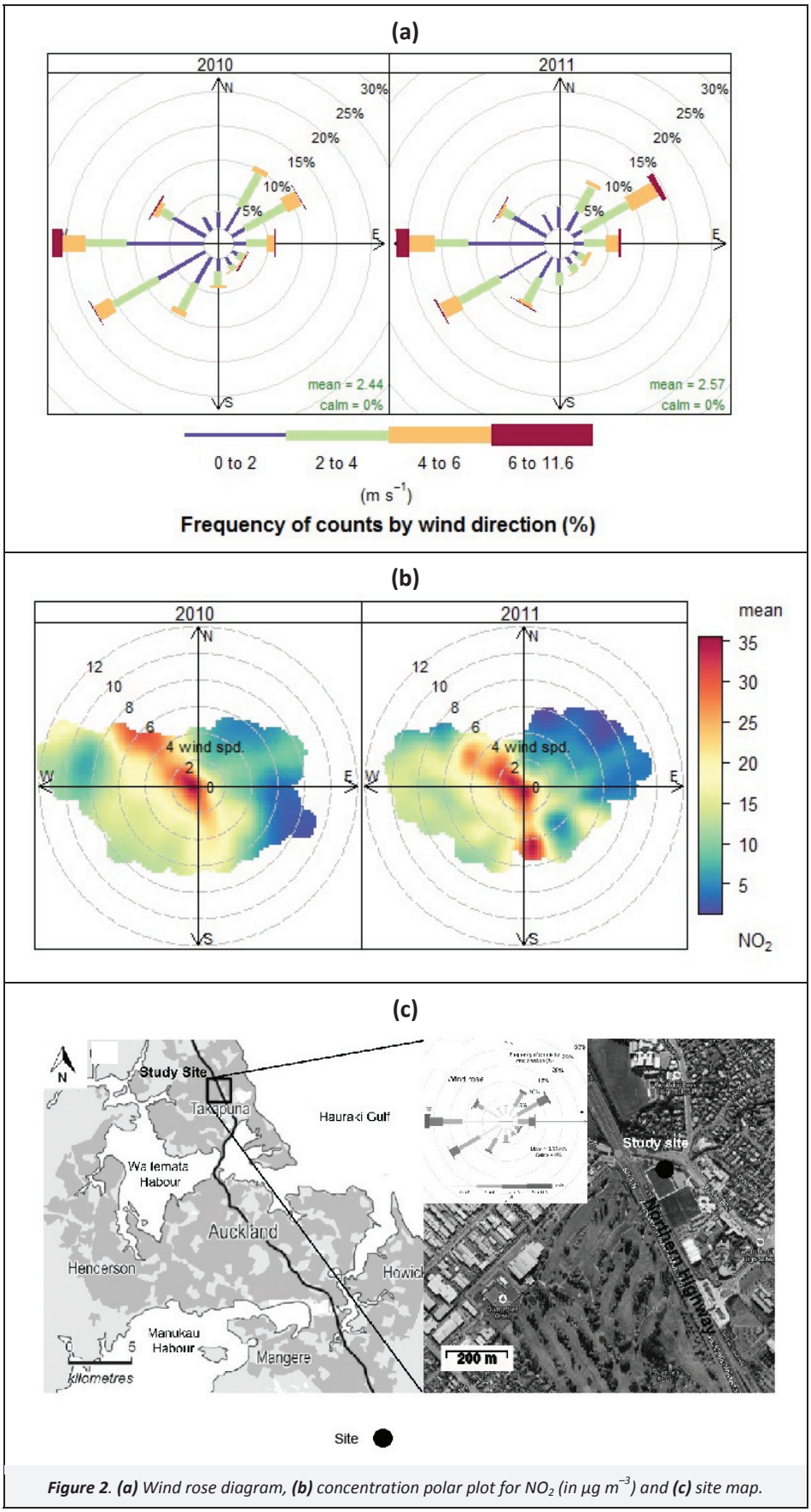

Figure 2. ***(a)*** *Wind rose diagram,* ***(b)*** *concentration polar plot for NO_2 (in µg m^{-3}) and* ***(c)*** *site map.*

Forward stepwise selection identified month and RH as insignificant inputs, while backward stepwise elimination gave its best network by eliminating only RH (Table 2). When genetic optimization is applied for the optimization of inputs, it considered that all predictor variables are important in making the best predictions (Table 2). Genetic optimization took the largest amount

of training time (Table 2) but provided the best prediction results (Table 3). Forward selection and backward elimination techniques resulted in models having marginal differences in performance statistics (Table 3). Therefore, the genetically–optimized network with all inputs with ten hidden neurons is considered for further analysis.

Table 1. *Statistical properties of input and output variables to the ANN model*

Parameter		Mean	Median	Minimum	Maximum
NO_2 (μg m^{-3})	2010	21.5	18.6	0.2	86.3
	2011	20.0	17.4	0	89.5
Wind speed (m s^{-1})	2010	2.4	2.2	0	11.6
	2011	2.6	2.3	0	10.1
Temperature (°C)	2010	16.1	16.2	3.8	28.0
	2011	16.1	16.1	2.6	27.0
Relative humidity (%)	2010	74.4	76.0	29.1	95.6
	2011	75.2	76.6	24.7	96.9
Solar radiation (W m^{-2})	2010	178.8	8.7	0	1 142.0
	2011	170.4	7.5	0	1 165.8

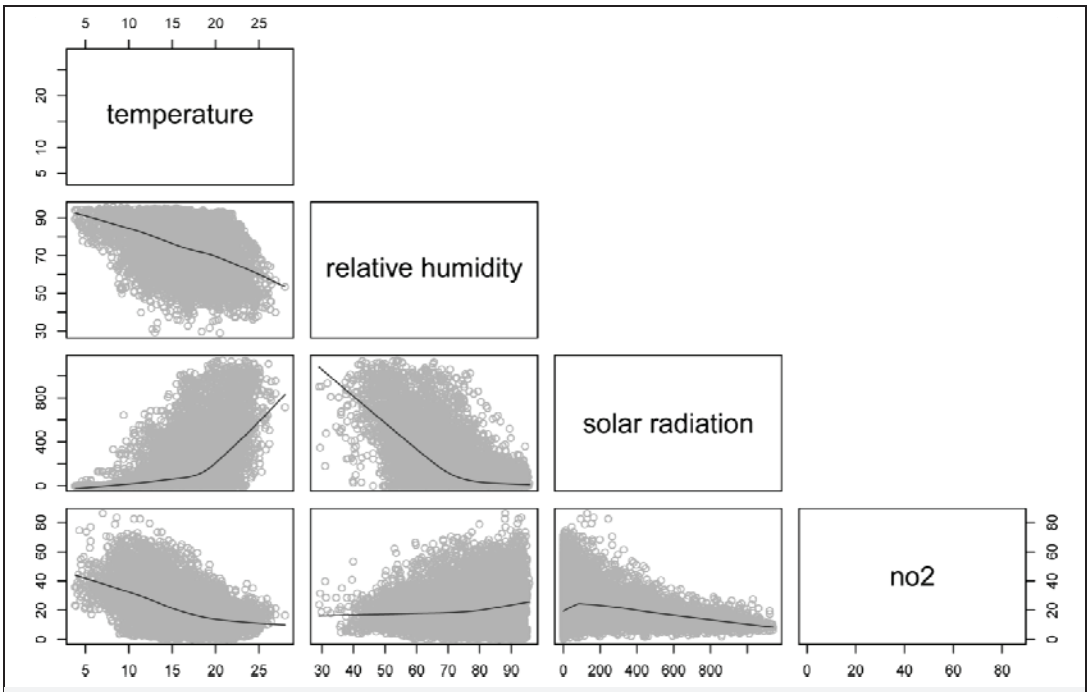

Figure 3*. Scatter plots showing how temperature, relative humidity and solar radiation are related to NO_2 concentrations and among each other. (The units; temperature: C, relative humidity: %, solar radiation: W m^{-2}, NO_2: μg m^{-3}).*

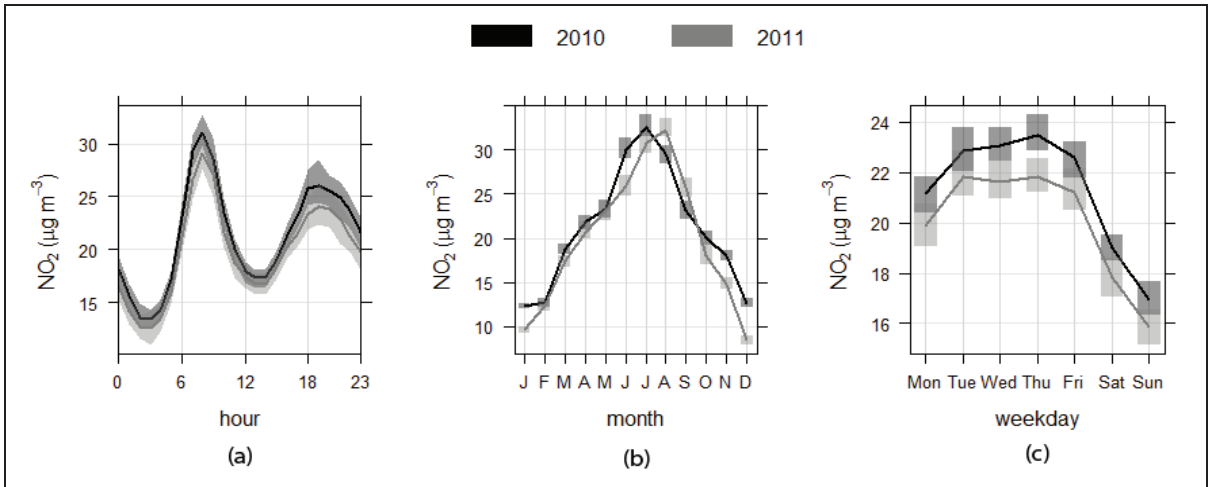

Figure 4*. **(a)** Hourly, **(b)** monthly, and **(c)** daily average variation of NO_2 concentration in 2010 and 2011 (Site: Takapuna, Auckland) (Note: The band in each plot shows 95% confidence interval of the mean).*

Table 2. *Summary results of three input optimization techniques*

Optimization Technique	Inputs Excluded	Number of Hidden Neurons	Approximate Training Time
Forward stepwise selection	Month, RH	37	37 min
Backward stepwise elimination	RH	37	25 min
Genetic Optimization	none	37	35 hours

Table 3. *Performance statistics of used optimization techniques*

Data Set	Statistical Parameter	Optimization Technique		
		Forward Selection	Backward Elimination	Genetic Optimization
Predicted for 2010	RMSE (μg m^{-3})	5.78	5.31	4.26
	IA	0.84	0.84	0.88
	r^2	0.86	0.90	0.93
Forecasted for 2011	RMSE (μg m^{-3})	7.57	7.50	4.79
	IA	0.79	0.79	7.07
	r^2	0.80	0.80	0.80

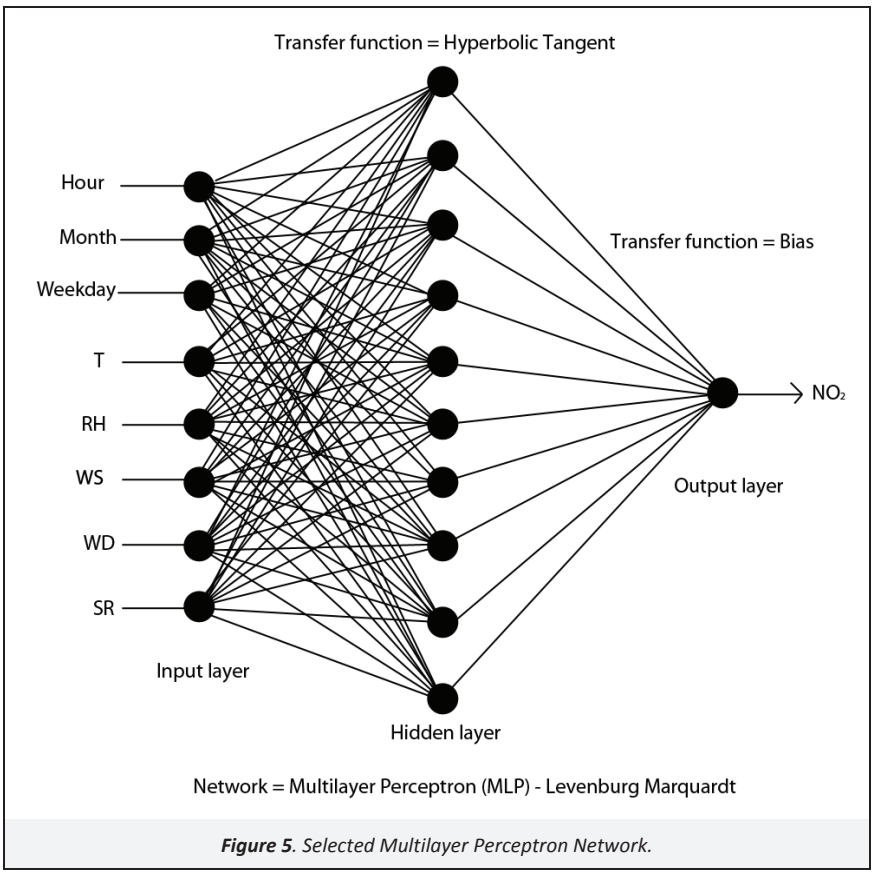

Figure 5*. Selected Multilayer Perceptron Network.*

3.4. Step III (b): Sensitivity analysis

The response of the above genetically–optimized ANN network to perturbations made to individual predictor variables in 100 small steps while locking in all other parameters to their mean value is presented in Figure 6. The responses of all eight predictor variables of the trained ANN network are obtained in this manner. The locked–in values of the predictor variables in each case are given in each plot to illustrate this fictitious situation. This analysis provides important information about the modeled ANN function, the sensitivity of each parameter and the non–linear relationship between each predictor variable and the modeled concentration. According to the modeled trends, the wind speed, wind direction and hour of the day show greatest sensitivity, both in terms of magnitude and also complexity, causing the concentrations to vary by approximately 40 μg m^{-3} over the observed range of the input parameter, while perturbations to all of the other variables, such as day of the week, month of the year, solar radiation and temperature result in only slight effects on the modeled concentration (a variation of approximately 10 μg m^{-3}). The relative humidity shows a flat response, indicating that it is the least significant predictor variable, consistent with the results of backward stepwise elimination.

For modeling NO_2 concentrations, amongst all of the parameters, the wind speed plays the most important role and shows a strong inverse relationship, consistent with observations (Figure 6a). It shows greater sensitivity and an almost linear relationship when wind speeds is greater than about 2 m s^{-1}. Perturbations in the wind direction, given in Figure 6b, reveal high concentrations when the winds are southerly (A) and north–westerly (B), in agreement with the concentration polar plots of actual observations (Figure 6b). This suggests that the model correctly captures the relationship between the NO_2 concentration

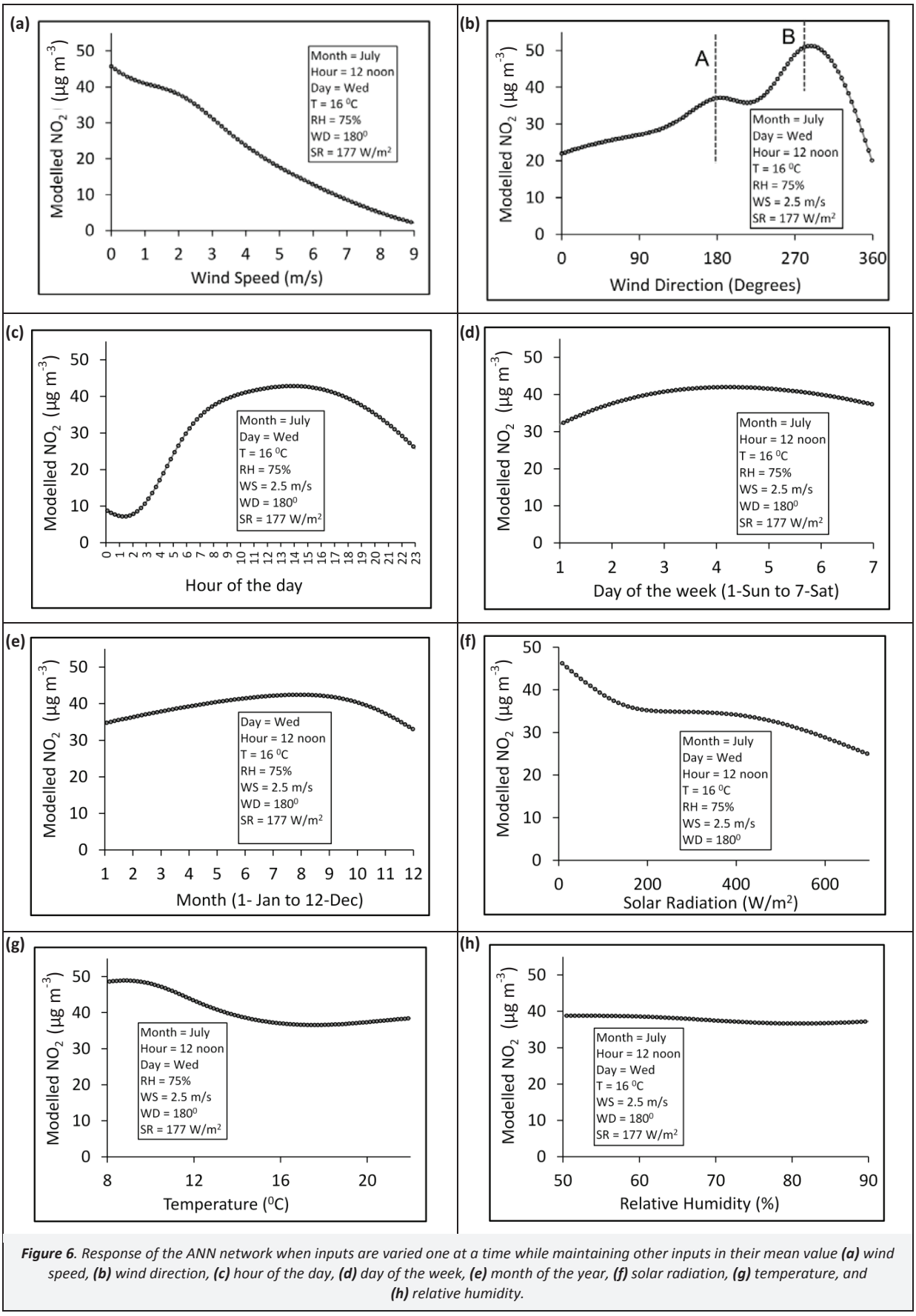

Figure 6*. Response of the ANN network when inputs are varied one at a time while maintaining other inputs in their mean value* ***(a)*** *wind speed,* ***(b)*** *wind direction,* ***(c)*** *hour of the day,* ***(d)*** *day of the week,* ***(e)*** *month of the year,* ***(f)*** *solar radiation,* ***(g)*** *temperature, and* ***(h)*** *relative humidity.*

and wind direction. The hour of the day input is expected to represent the sensitivity of the emission rate to NO_2 concentrations. Since all other parameters, including wind speed/wind direction, are locked in to their mean values, this response curve can be expected to represent the individual effect of the hourly variation in the NO_2 emission rate. Perturbations to day of the

week revealed high concentrations during weekdays and low concentrations during weekends with minimum concentrations on Sunday (Figure 6d). This behavior is in agreement with what is expected at this site (Figure 4c). Sensitivity analysis in relation to month of the year revealed slightly elevated concentrations during the winter months (Figure 6e), as expected in Figure 4b. The modeled NO_2 concentrations decrease slightly with increasing temperature and solar radiation (Figure 6f and 6g). This is expected behavior (night–time low temperature, low solar radiation, low concentration) while no changes in NO_2 concentrations are observed with perturbations in the relative humidity (Figure 6h). A major question about sensitivity analysis raised in the literature is whether all combinations of these modeled fictitious situations adequately represent reality (Lek et al., 1996). For example, a mean temperature of 16 °C is not present throughout the day or throughout the year and the solar radiation intensity is not the same throughout the day. However, this comparison of modeled trends through sensitivity analysis with actual trends presented in Section 3.1 revealed that sensitivity analysis can be relied upon. The extensive analysis provided in this study further strengthens the work by Yan Chan and Jian (2013).

3.5. Step III (c): ANN networks with eliminated inputs

The list of models developed by successively removing the least significant inputs is given in Table 4. Each model is subjected to step size, momentum and processing element optimization based on a genetic algorithm to ensure the best network parameters are achieved in each case. Model 1 to Model 4 are created by the successive removal of the relative humidity, temperature and solar radiation until Model 4 has wind speed and wind direction as the only meteorological parameters. Model 5 is created by eliminating wind speed and wind direction while retaining the other six predictor variables, and Model 6 is created by eliminating month, hour and day inputs while retaining all five of the meteorological parameters as predictor variables. The latter two models assess the collective sensitivity of wind speed/wind direction and hour of the day, month of year and day of week inputs on NO_2 concentrations.

Table 4. *Input used in models for predicting and forecasting NO_2 concentration*

Model Name	Inputs
Model 1	Month, Hour, Day, Temp, RH, WS, WD, SR
Model 2	Month, Hour, Day, Temp, WS, WD, SR
Model 3	Month, Hour, Day, WS, WD, SR
Model 4	Month, Hour, Day, WS, WD
Model 5	Month, Hour, Day, Temp, RH, SR
Model 6	Temp, RH, WS, WD, SR

The performance statistics of the model based on the whole test data set (for the whole of 2011) is given in Table 5. The results suggest that the best model accuracy is achieved by Model 1 that uses all eight predictor variables. However, the successive removal of relative humidity, temperature and solar radiation (Model 2, Model 3 and Model 4) has only a very minimal effect on model statistics. This is in agreement with the results of the sensitivity analysis that showed that the modeled concentration is least sensitive to these parameters. When the wind speed and wind direction are removed from the network (Model 5), a drastic reduction in model performance is observed, suggesting that they are the most significant predictors in this model. Similarly, when hour, day and month inputs are not used in the network (Model 6), the network did not perform well. The time series of observed versus predicted concentrations of the best model (Model 1) for predicting for summer and winter 2010 and forecasting for summer and winter 2011 are given in Figure 7a and 7b, respectively. It shows that the model correctly captures the trends where concentrations are low during the summer and high during the winter. The forecasting accuracy of the model for the following year is satisfactory.

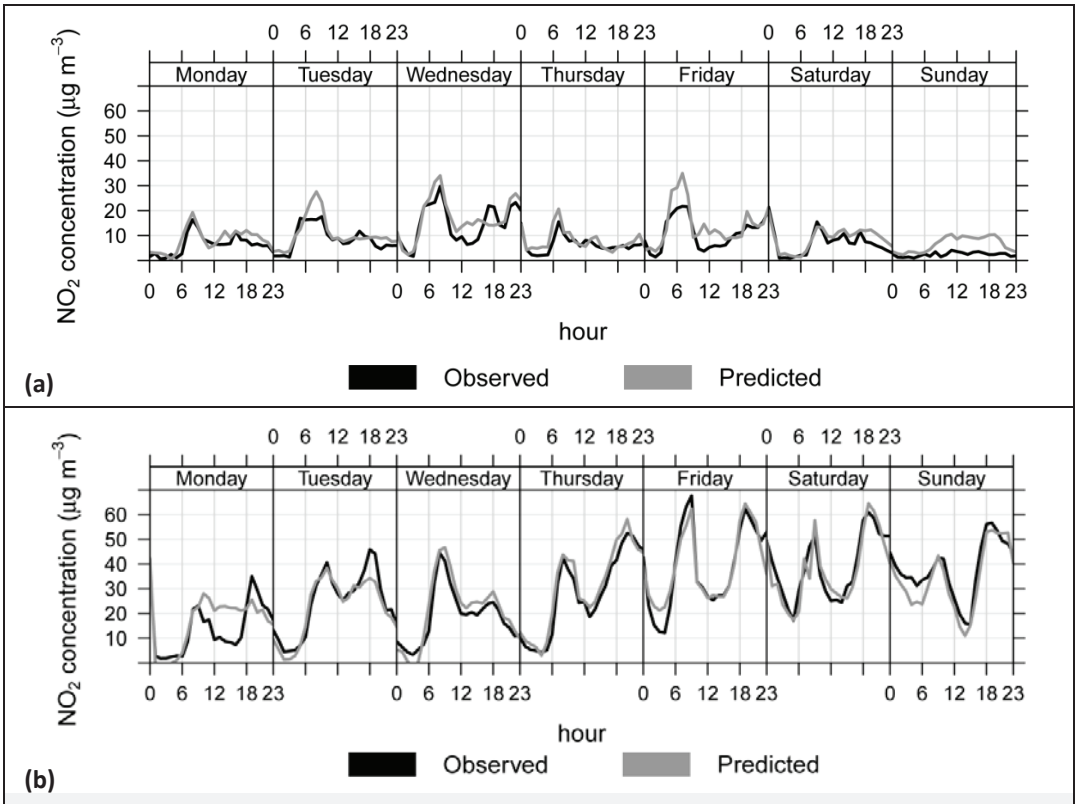

Figure 7*. Predictions and forecasts made for four test weeks using Model 1.* ***(a)*** *Predicted for summer 2010,* ***(b)*** *predicted for winter 2010,* ***(c)*** *Forecasted for summer 2011,* ***(d)*** *Forecasted for winter 2011.*

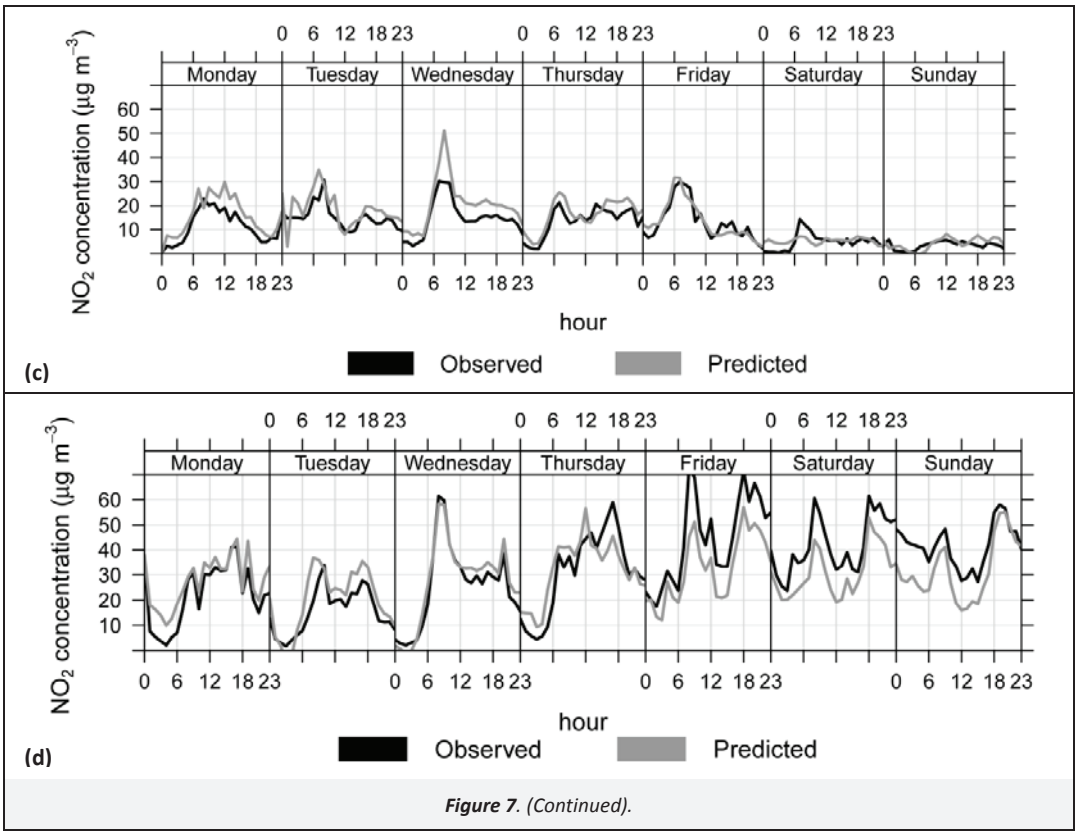

***Figure 7**. (Continued).*

***Table 5.** Performance statistics of the developed ANN models*

Data Set	Statistical Parameter	Model Name					
		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Predicted for 2010	MB (µg m^{-3})	0.93	0.83	1.10	0.79	3.18	2.75
	RMSE (µg m^{-3})	4.26	4.82	4.60	4.63	11.32	7.97
	IA	0.88	0.86	0.87	0.87	0.66	0.78
	r^2	0.93	0.92	0.92	0.92	0.55	0.77
	FAC2	0.85	0.84	0.87	0.87	0.59	0.74
Forecasted for 2011	MB (µg m^{-3})	1.01	1.08	0.98	1.01	1.81	1.51
	RMSE (µg m^{-3})	7.07	6.90	7.13	7.07	10.83	9.64
	IA	0.79	0.79	0.79	0.79	0.65	0.69
	r^2	0.80	0.77	0.77	0.77	0.48	0.58
	FAC2	0.87	0.88	0.87	0.88	0.73	0.78

A comparison in performance statistics of Model 1 and Model 4, separated according to seasons, is presented using a Taylor diagram in Figure 8. Both models perform best in autumn, suggesting that concentrations are closely related to wind speed and direction and that emission rates are well represented by the hour, weekday and month inputs. The RMSE and correlation coefficients of both models are similar in summer and autumn, but the modeled concentrations using both models show different standard deviations compared to observations. They perform worst in spring with respect to correlation and RMSE statistics but the standard deviations of the modeled values are similar to those of the observed. This analysis shows that Model 1 and Model 4 do not perform significantly differently across the year, despite seasonal effects affecting emission rates and wind speed and wind direction interactions.

3.6. Step IV: Analysis of the results of Step I and Step II

- Foreword selection, backward elimination, sensitivity analysis and the iterative procedure identified that relative humidity is the least significant predictor variable; the inclusion of relative humidity only marginally improves model performance.
- Sensitivity analysis and the iterative procedure identified that temperature and solar radiation contribute little to the variability in NO_2. However, the inclusion of temperature and solar radiation as predictor variables marginally improves model performance.
- Results of the sensitivity analysis match with the expected trend associated with the observed data, hence the optimized ANN function can be trusted for "what–if scenario analysis".
- Sensitivity analysis and comparison of Model 1 and Model 5 reveals that wind speed and wind direction are the most sensitive inputs.
- Sensitivity analysis and comparison of Model 1 and Model 6 reveals that the predictor variables hour, weekday and month provide the information needed by the model to account for variations in emission rates according to these time scales.
- While the model with five meteorological parameters could explain 80% of the variability in the NO_2 concentration of the following year, the model with only wind speed and wind direction could explain 77% of the variability. This finding supports the results of a previous study using a semi–empirical box model (SOSE) for predicting site–specific concentrations of pollutants in New Zealand using only wind speed and direction data (Dirks et al., 2002; Dirks et al., 2003).

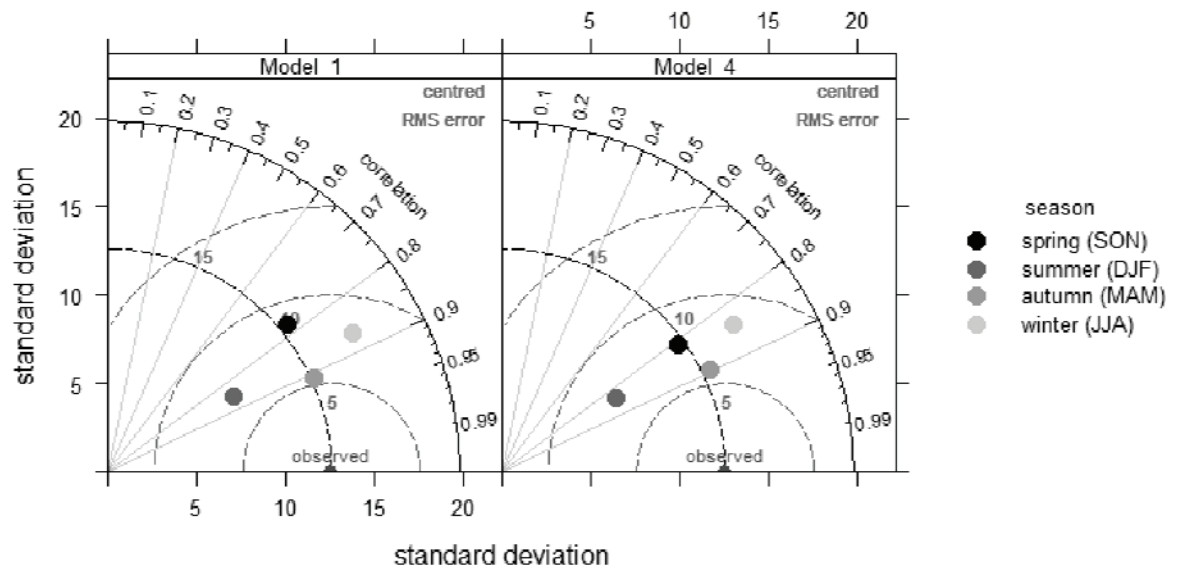

***Figure 8**. Performance comparison of Model 1 and Model 4 according to seasons of the year (Note: A Taylor Diagram compares the statistical performance of a model using three performance statistics simultaneously (Taylor, 2001). The statistics are the correlation coefficient; standard deviation and the (centered) root mean squared error (RMSE). Since these statistics are related by the law of cosines they can be represented in a 2D diagram. The Standard deviation of the observed values is marked on the x–axis using a half circle. The standard deviation of each model should be read as the radial distance from the origin of the plot (0, 0). Therefore if the model predictions have similar standard deviation to the observed values it should be marked closer to the black dashed radial line marked as the standard deviation of the observed values. The correlation coefficients are shown on the arc while the values closer to the x–axis are closer to one. Correlation coefficient of a specific model should be read along the radially drawn straight line through the origin. The RMSE is centered by subtracting the mean values of the observed and predicted values before calculating the RMSE for it to be marked as concentric gray dashed lines emanating from the "observed" point. Therefore the best model should lie as much as close to the "observed" point with regards to standard deviation, correlation coefficient and RMSE).*

3.7. Comparison with multi linear regression (MLR) models

To compare the ANN model performance with linear approaches used to develop a relationship between concentration, meteorology and time parameters, three MLR models were developed with different input parameters. Similar to the ANN models, the MLR models are trained on the 2010 data set and predicted on the 2011 data set of the same site. Input parameters with the functions of the three different models are:

$$MLR\_1 = 0.293855(month) + 0.318272(hour) + 0.346191(day) + 14.3482 \quad (6)$$

$$MLR\_2 = 0.354404(month) + 0.598358(hour) + 0.311219(day) - 4.16081(wind\ speed) + 0.062221(wind\ direction) + 9.226765 \quad (7)$$

$$MLR\_3 = -0.12103(month) + 0.710208(hour) + 0.287667(day) - 1.49361(temperature) 0.07516(RH) - 3.70301(wind\ speed) + 0.047006(wind\ direction) + 0.012869(Solar\ radiation) + 29.07641 \quad (8)$$

The model performance statistics of the three models compared to that of the simplified ANN model (Model 4) are given in Table 6. This comparison is evidence that even the simplified ANN model is capturing non–linearity in the relationship between meteorology and observed NO_2 concentrations that a linear model is unable to capture. Figure 9 shows a comparison between the simplified ANN model and the best MLR model in capturing the average behavior of the time series. The shaded areas represent the 95% confident interval of the respective time series. According to this comparison, the ANN model is showing very good agreement with the hourly variation of the observed concentrations and becomes the best tool for modeling the time series of NO_2 out of the models considered in this study.

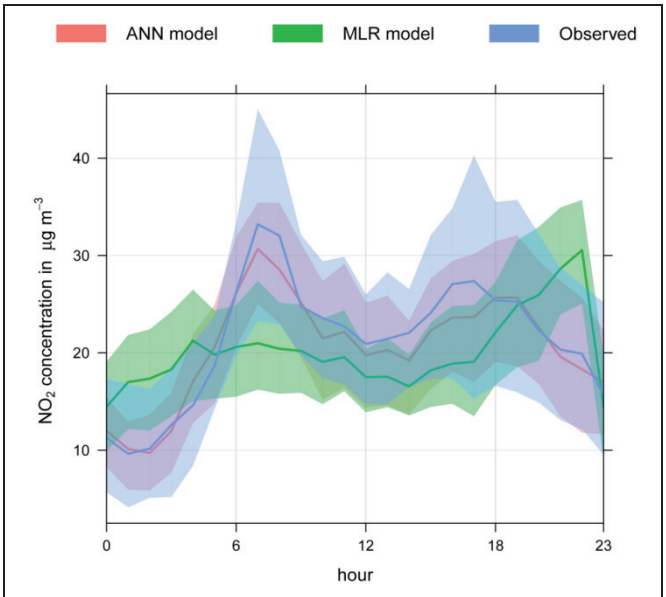

***Figure 9.** Comparison of performance of the best MLR model with the simplified ANN model in capturing the hourly time variation of NO_2 concentration.*

4. Conclusions

This study reveals that, by carefully choosing inputs to represent monthly, daily and hourly emission patterns and relationships to wind speed, wind direction, atmospheric temperature, relative humidity and solar radiation, a simple ANN model can give reliable forecasts of nitrogen dioxide concentrations. These

predictor variables are easily obtainable from routine monitoring networks and hence the developed model is practically applicable to many situations. Despite the fact that NO_2 is a reactive species, surface winds are found to be the most influencing parameter in determining the ambient NO_2 concentrations in Auckland, New Zealand and an ANN model that uses only surface wind speed and direction as the meteorological parameter can reliably forecast NO_2 concentrations for the subsequent years. Once most influencing factors have been identified through sensitivity analysis, even a single network can be used for multiple pollutants based on the common influencing parameters. A similar methodology could be applied to other scenarios in which the meteorology and emission patterns influence pollution concentrations, albeit in different ways, such as for different road emission scenarios or for industrial sources.

Table 6. *Performance statistics of MLR models compared to ANN model*

Model	MB ($\mu g\ m^{-3}$)	RMSE ($\mu g\ m^{-3}$)	IA	r^2	FAC2
MLR–1	–0.58	15.79	0.53	0.17	0.54
MLR–2	–0.72	13.73	0.59	0.32	0.63
MLR–3	–1.16	10.82	0.70	0.60	0.74
ANN (Model 4)	1.01	7.07	0.79	0.77	0.88

The presented model has been tested with large blocks of data removed from the training time series (in this case, two weeks from summer and winter from the 2010 data set) and the model could still reliably forecast for the same weeks of the consecutive year. This is an added advantage compared with time–lagged models.

The sensitivity analysis of inputs has proven to be a useful technique for understanding the mechanism of the modeled function, giving more insight into the internal mechanism of the ANN model. It also helps to understand the relative contribution of input parameters that can be used to eliminate irrelevant inputs to make the model more robust. If the sensitivity analysis of the input variables gives the expected trends in real measurements, it guarantees that the ANN model has captured the governing dispersion mechanism specific to the site. Hence, the developed ANN model can be used for "what–if scenario analysis"; if the emission rates are included as a model input parameter, different emission scenarios can be analyzed separately from the meteorological scenarios. This would be a valuable future exercise. The authors believe that the inclusion of emission rates would further improve model performance and that these predictor variables would help the model to capture the complexities related to emissions that have not been captured in the present model. For example, uncertainties of the present model caused by any changes in emission rates from year 2010 to year 2011 could be eliminated by the introduction of emission rates as an input to the model. The bivariate polar plots presented reveal that there are many wind speed/wind direction/temperature clusters present and the inclusion of these clusters as an input could help to fine–tune the model. Hence, it would be valuable to investigate the inclusion of clusters (for example, using the k–means clustering of temperature/winds/concentration) into the model (Carslaw and Ropkins, 2012).

Acknowledgments

The authors would like to thank Auckland Council for providing air pollution and meteorological data for the study.

References

Antanasijevic, D.Z., Pocajt, V.V., Povrenovic, D.S., Ristic, M.D., Peric–Grujic, A.A., 2013. PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization. *Science of the Total Environment* 443, 511–519.

Arhami, M., Kamali, N., Rajabi, M.M., 2013. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by monte carlo simulations. *Environmental Science and Pollution Research* 20, 4777–4789.

Burnett, R.T., Cakmak, S., Brook, J.R., 1998. The effect of the urban ambient air pollution mix on daily mortality rates in 11 Canadian cities. *Canadian Journal of Public Health–Revue Canadienne De Sante Publique* 89, 152–156.

Carslaw, D.C., 2014. The openair manual — open-source tools for analyzing air pollution data. Manual for version 1.0, King's College London.

Carslaw, D.C., Ropkins, K., 2012. Openair – an R package for air quality data analysis. *Environmental Modelling & Software* 27–28, 52–61.

Carslaw, D.C., Beevers, S.D., Tate, J.E., Westmoreland, E.J., Williams, M.L., 2011. Recent evidence concerning higher NO_X emissions from passenger cars and light duty vehicles. *Atmospheric Environment* 45, 7053–7063.

Carslaw, D.C., Carslaw, N., 2007. Detecting and characterising small changes in urban nitrogen dioxide concentrations. *Atmospheric Environment* 41, 4723–4733.

Chelani, A.B., Rao, C.V.C., Phadke, K.M., Hasan, M.Z., 2002. Prediction of sulphur dioxide concentration using artificial neural networks. *Environmental Modelling & Software* 17, 161–168.

Cheng, S.Y., Li, L., Chen, D.S., Li, J.B., 2012. A neural network based ensemble approach for improving the accuracy of meteorological fields used for regional air quality modeling. *Journal of Environmental Management* 112, 404–414.

Davy, P., Trompetter, B., Markwitz, A., 2009. Source apportionment of airborne particles in the Auckland region: 2008 Update. *GNS Science Consultancy Report 2009/165,* Auckland, GNS.

Dirks, K.N., Johns, M.D., Hay, J.E., Sturman, A.P., 2003. A semi–empirical model for predicting the effect of changes in traffic flow patterns on carbon monoxide concentrations. *Atmospheric Environment* 37, 2719–2724.

Dirks, K.N., Johns, M.D., Hay, J.E., Sturman, A.P., 2002. A simple semi–empirical model for predicting missing carbon monoxide concentrations. *Atmospheric Environment* 36, 5953–5959.

Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NO_X and NO_2 concentrations in urban air in London. *Atmospheric Environment* 33, 709–719.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences. *Atmospheric Environment* 32, 2627–2636.

Grivas, G., Chaloulakou, A., 2006. Artificial neural network models for prediction of PM_{10} hourly concentrations, in the greater area of Athens, Greece. *Atmospheric Environment* 40, 1216–1229.

Hewson, E. W., 1956. Meteorological factors affecting causes and controls of air pollution. *Journal of the Air Pollution Control Association* 5, 235–241.

Hrust, L., Klaic, Z.B., Krizan, J., Antonic, O., Hercog, P., 2009. Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. *Atmospheric Environment* 43, 5588–5596.

Jiang, D.H., Zhang, Y., Hu, X., Zeng, Y., Tan, H.G., Shao, D.M., 2004. Progress in developing an ANN model for air pollution index forecast. *Atmospheric Environment* 38, 7055–7064.

Jiang, N., Hay, J.E., Fisher, G.W., 2005. Effects of meteorological conditions on concentrations of nitrogen oxides in Auckland. *Weather and Climate* 24, 15–34.

Karatzas, K.D., Kaltsatos, S., 2007. Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece. *Simulation Modelling Practice and Theory* 15, 1310–1319.

Kurt, A., Oktay, A.B., 2010. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Systems with Applications* 37, 7986–7992.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996. Role of some environmental variables in trout abundance models using neural networks. *Aquatic Living Resources* 9, 23–29.

Malby, A.R., Whyatt, J.D., Timmis, R.J., 2013. Conditional extraction of air–pollutant source signals from air–quality monitoring. *Atmospheric Environment* 74, 112–122.

Nagendra, S.M.S., Khare, M., 2006. Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. *Ecological Modelling* 190, 99–115.

Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* 178, 389–397.

Olden, J. D., Jackson, D.A., 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* 154, 135–150.

Perez, P., 2012. Combined model for PM_{10} forecasting in a large city. *Atmospheric Environment* 60, 271–276.

Principe, J. C., Euliano, N.R., Lefebvre, W.C., 1999. *Neural and Adaptive Systems*: *Fundamentals Through Simulations* with CD–ROM, John Wiley & Sons Inc.

Samarasinghe, S., 2006. *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition*, CRC Press.

Singh, K.P., Gupta, S., Kumar, A., Shukla, S.P., 2012. Linear and nonlinear modeling approaches for urban air quality prediction. *Science of the Total Environment* 426, 244–255.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research–Atmospheres* 106, 7183–7192.

Yan Chan, K.Y., Jian, L., 2013. Identification of significant factors for air pollution levels using a neural network based knowledge discovery system. *Neurocomputing* 99, 564–569.

Willmott, C.J., Robeson, S.M., Matsuura, K., 2012. A refined index of model performance. *International Journal of Climatology* 32, 2088–2094.